Use of Discontinuous Methods of Data Collection in Behavioral Intervention: Guidelines for Practitioners

Kate Fiske and Lara Delmolino
Douglass Developmental Disabilities Center
and Rutgers, The State University of New Jersey

ABSTRACT

Over the past three decades, researchers have examined the sensitivity and accuracy of discontinuous data-collection methods. Momentary-time sampling (MTS) and partial-interval recording (PIR) have received particular attention in regards to their ability to estimate the occurrence of behavior and their sensitivity to behavior change compared to continuous data collection. In this article, we summarize these findings and provide recommendations for designing a discontinuous measurement system with consideration of the dimensions of behavior to be measured and the expected direction of behavior change.
Keywords: discontinuous data collection, momentary time sampling, partial interval recording

Direct observation and measurement of behavior is a defining characteristic of applied behavior analysis as a science and as a practice (Baer, Wolf, & Risley, 1968). The data generated via direct observation serve as the basis upon which practitioners make treatment decisions and evaluate treatment effects. Therefore, designing a high-quality measurement system is an essential early step in developing and evaluating a behavioral intervention. The "quality" of a direct measurement system is determined by the extent to which the system (a) generates accurate data, (b) produces reliable outcomes, and (c) is sensitive to changes in the occurrence of behavior. Each of these features will be determined in part by the operational definitions written to guide data collectors and by the training level of the data collectors implementing those systems. Quality also is influenced considerably by the specifics of the data collection system itself.

Continuous data collection systems are those that capture every possible behavioral occurrence, either by recording each instance of a behavior (i.e., frequency recording) or by recording the number of seconds each instance of behavior occurs (i.e., duration recording) during an observation. These systems offer a complete record that can be reported in standard scientific units, such as responses per minute or percentage of observation (Johnston & Pennypacker, 2009). Alternatively, discontinuous data collection systems are those that capture only a sample of behavior during an observation. These systems involve dividing an observation into equal duration intervals, and scoring the occurrence or nonoccurrence of behavior within each interval (Mudford, Taylor, & Martin, 2009). Three variants of discontinuous measurement have been described in the literature; these systems differ in terms of how they define—and how an observer scores—a behavioral occurrence or nonoccurrence during each interval. When using partial-interval recording (PIR), an occurrence is defined as an instance of target behavior that occurs at any time during the interval. Whole-interval recording (WIR) defines an occurrence as when the target behavior occurs for the entire duration of the interval. Momentary-time sampling (MTS) defines an occurrence only if the target behavior occurs as the interval ends, frequently in the last second of the interval. Data are reported as the percentage of intervals during which behavior was scored.

Due in large part to their ease of implementation, discontinuous measurement systems are popular in practice and in applied research. For example, Mudford, Taylor, and Martin (2009) found that 45% of studies published in the *Journal of Applied Behavior Analysis* from 1995 to 2005 used discontinuous measures of data collection. Given that research protocols are often characterized by more rigorous measurement (such as continuous data collection) than everyday behavioral practice, this number is likely an underestimate of the extent to which discontinuous data collection methods are used in clinical and educational settings. Despite the widespread use of discontinuous measurement, clear recommendations for designing an optimal measurement system, with consideration of factors affecting measurement accuracy, have not been presented in a consolidated format. The purpose of this paper is to provide recommendations on designing measurement systems based upon the published literature with regard for such variables as dimensions of behavior, interval duration, and session duration. We have organized this paper in terms of the considerations a practitioner would make when designing a measurement system.

What measurement error is introduced by discontinuous measurement?

Because discontinuous systems involve taking only a sample of ongoing behavior, concerns about measurement accuracy and sensitivity are inherent in discontinuous data. Due to the specific scoring rules associated with each discontinuous measurement system, the directionality of measurement error is often consistent and predictable. PIR will consistently overestimate the true occurrence of behavior, since both a 1-s response and a 9-s response in each 10-s interval are coded identically as an *occurrence*. WIR will consistently underestimate the true occurrence of a behavior, with both a 1-s response and a 9-s response coded identically as a *nonoccurrence*. MTS is not associated with a characteristic direction of error. Because of this fact, MTS will most commonly provide a more accurate estimate of behavioral duration than either PIR or WIR, although the accuracy will be impacted by variables such as behavior duration or level (Ciotti Gardenier, MacDonald, & Green, 2004; Powell, Martindale, & Kulp, 1975; Powell, Martindale, Kulp, Martindale, & Bauman, 1977).

When should I collect discontinuous data instead of continuous data?

Continuous data collection methods are not associated with the characteristic measurement error inherent in discontinuous methods. Continuous methods, therefore, offer a more accurate (i.e., higher quality) measurement of the target behavior and are preferred relative to discontinuous methods under most circumstances. However, the accuracy and feasibility of continuous methods may be impacted by the specific clinical situation. Discrete behavior, in which the onset and offset of behavior are clear and easily recorded (e.g., punches) is more easily captured via continuous methods than is behavior with more ambiguous breaks between instances (e.g., vocal behavior in which there may be brief pauses between utterances). Further, the amount of attention or effort required by the observer may also impact the accuracy of continuous measurement. For instance, counting each instance of a behavior that occurs at very high rates (e.g., words uttered in a conversation or the number of hand flaps by a child engaging in stereotypy), recording multiple responses simultaneously, or collecting data while engaging in other activities (e.g., a when a teacher is recruited to collect data during their school day) may all adversely affect the accuracy of continuous measurement. While this makes intuitive sense, the impact of these factors on accuracy has not been definitively examined in research. In light of these factors, however, practitioners may choose discontinuous systems (which tend to be easier to use) under conditions in which there are multiple demands upon data collectors. For the remainder of this review, we will assume the practitioner has decided discontinuous systems were appropriate given the exigencies of their case. We will address considerations in designing a measurement format in a series of questions the behavior analyst should ask in this process.

Which discontinuous system should I use?

Behavior analysts are typically called upon to intervene upon behavioral deficits (i.e., increasing low-occurrence behavior) and behavioral excesses (i.e., decreasing high-occurrence behavior). These decisions impact the selection of a measurement system. For instance, in increasing a low-occurrence behavior, the overestimation characteristic of PIR will inflate the true levels of behavior during baseline and generate "ceiling" effects that may be insensitive to increases in behavior following an intervention (e.g., the increase in behavior from 2-s each interval to 10-s each interval would be masked) (Harrop & Daniels, 1986; Rapp, Colby-Dirksen, Michalski, Carroll, & Lindenberg, 2008). Perhaps more important is that PIR's overestimation may give the impression that behavior is occurring at a higher level following intervention than it is in reality. In an extreme case, imagine the outcome of an intervention designed to increase the on-task behavior of a child in a classroom in which, following an intervention, the child is on task only 3 s during each 10-s interval. Using PIR, the graphic depiction of this case would falsely indicate that treatment was highly effective (100% on-task), when the true value would be much lower (30% on-task). For these reasons, PIR is not recommended when treating behavioral deficits in which it is important for behavior to occur at very high levels (e.g., on-task behavior).

On the other hand, WIR also provides some interpretive difficulties. The characteristic underestimation would require a very robust behavior change in order to show an increase in behavior (i.e., an increase from a 1-s to a 9-s occurrence in each 10-s interval would not be captured). Thus, WIR may be an appropriate, but overly stringent measurement system in many situations. MTS has generally been shown to be the most accurate measurement system and is recommended for measurement in cases of treating behavioral deficits. However, practitioners should be aware that MTS may have difficulty capturing low-frequency behavior (e.g., Saudargas & Zanolli, 1990) at baseline and may be insensitive to small changes following the initiation of treatment.

Similar considerations arise when treating behavioral excesses. WIR will underestimate the true occurrence of behavior and therefore is likely to underestimate the level of the behavior before treatment. Additionally, the underestimation produced by WIR is likely to depict robust treatment effects even in cases in which behavior persisted (e.g., a behavior that occurred for 9 s during each 10-s interval would be consistently scored as a nonoccurrence). For this reason, WIR is not recommended in cases of treating behavioral excesses. On the flip-side, PIR will overestimate the occurrence of behavior (both during baseline and following treatment), and the ceiling effect may make it difficult to capture small to moderate change in behavior. If a high-rate behavior targeted for decrease is measured at a ceiling of 100% at baseline, a small decrease in behavior may not be captured initially using PIR, as it may still overestimate that the behavior occurs 100% of the session. Additionally, PIR will require near complete elimination of behavior in order to demonstrate a clinically significant treatment effect. In this regard, PIR data collection is appropriate for treatment of behavioral excesses, but will be very stringent. MTS would likely provide the most accurate estimate of behavior; however this accuracy may be somewhat compromised depending on the effectiveness of the intervention. That is, if behavior persists at infrequent levels with brief durations, these instances may be missed by MTS (Saudargas & Zanolli, 1990). Selection of a measurement system in the case of reducing behavioral excesses may be dictated in part by the necessity that behavior be eliminated to zero levels (e.g., eye gouging); this could only be captured accurately via PIR. However, if zero levels of behavior are not essential, MTS would be appropriate.

For what duration should each interval last?

Simply put, the briefer the interval duration, the less systematic error will be introduced into a measurement system, regardless of whether PIR, WIR, or MTS are implemented (Devine, Rapp, Testa, Henrickson & Schnerch, 2011; Hanley, Cammilleri, Tiger, & Ingvarsson, 2007; Powell et al., 1975; Powell et al., 1977; Rapp et al., 2008). However, the decrease in interval duration requires additional effort from data collectors and therefore increases the amount of human error that is introduced during measurement. For instance, Hanley et al. compared the interobserver agreement of two trained data collectors using an MTS system when interobservation intervals ranged from between 1.5 to 6.0 s. Observers reliably coded a measure of preschoolers' engagement with intervals as brief as 3 s, but interobserver agreement fell to unacceptable levels when intervals were less than 3 s. Again, the frequency with which data collectors can score intervals accurately may also be related to the number of behaviors they are scoring simultaneously and the number of other simultaneous responses they are required to emit (e.g., caregiving). It is also worth noting that as intervals approach 1 s, the data collection system then approximates continuous measurement and thus calls into question whether continuous measurement would be possible.

It is difficult to specify the upper-bound at which each data collection system can extend and it likely differs between PIR, WIR, and MTS. The most common interval in the available literature implementing these systems is 10 s, but in most cases 10 s intervals with PIR will still overestimate behavior. MTS tends to maintain greater accuracy than PIR at longer interval durations. For instance, Hanley et al. (2007) found less than 5% difference in behavioral estimates when MTS intervals were between 5 s and 120 s. Research has not fully articulated the conditions under which longer duration intervals will continue to accurately capture behavior for each system, but is likely related to the variability in behavior within sessions and the duration of observations.

How long should each observation last?

Intuitively, longer sample observations reduce the amount of error in the measurement system, as a longer sample is more representative of the whole time of interest (e.g., the full session or day the observation sample is to represent). However, different factors will influence the representativeness of the sample. Mudford and Beale (1990) compared the accuracy of observation samples of various lengths to a full 2.5 hr session when collecting continuous data. The authors found that a longer observation sample (i.e., 105 min) was needed to more accurately represent a low-occurrence behavior during the full session, whereas a shorter interval (i.e., 30 min) accurately represented a high-occurrence behavior. A similar investigation by Devine et al. (2011) examined the impact of session length on the accurate detection of functional control by PIR and MTS using simulated data. When comparing discontinuous data to continuous data, the authors found, as one example, that the use of 30-s MTS across a 30-min session (or longer) was accurate and sensitive when evaluating behavior change. However, much more research needs to be done before definitive recommendations can be made. For now, practitioners should err on the side of caution and use longer observations when possible, especially when measuring low-frequency and/or high-variability behavior.

Would my treatment decisions and interpretations be different based upon the measurement system I choose?

Several studies have examined differences in how data would be interpreted based upon the measurement system implemented. For example, Meany-Daboul, Roscoe, Bourret, and Ahearn (2007) considered five treatment data sets in which they demonstrated clear reductions in the occurrence of problem behavior, recorded using continuous methods, in an ABAB experimental design. The experimenters then reconstructed their figures as they would have appeared had they recorded behavior using 10-s PIR and 10-s MTS and compared the extent to which an expert panel and a structured criteria would differ in their nominations of a demonstration of functional control. The authors reported strong correspondence between ratings of functional control across each data collection method, suggesting that the amount of error introduced (at least when recording via 10-s PIR and MTS) was not sufficient to change ratings of an experimental effect. The authors did report 10-s PIR to be slightly more sensitive when detecting change in behaviors initially scored using continuous frequency recording and 10-s MTS to be slightly more sensitive in detecting change in behaviors scored using continuous duration recording. It is also worth noting that each of these instances were cases of behaviors targeted for reduction and although the behavior change remained apparent following their re-analyses with discontinuous systems, there were concerns regarding the clinical significance of the behavior change. That is, although a behavior change was apparent, the levels of behavior appeared higher

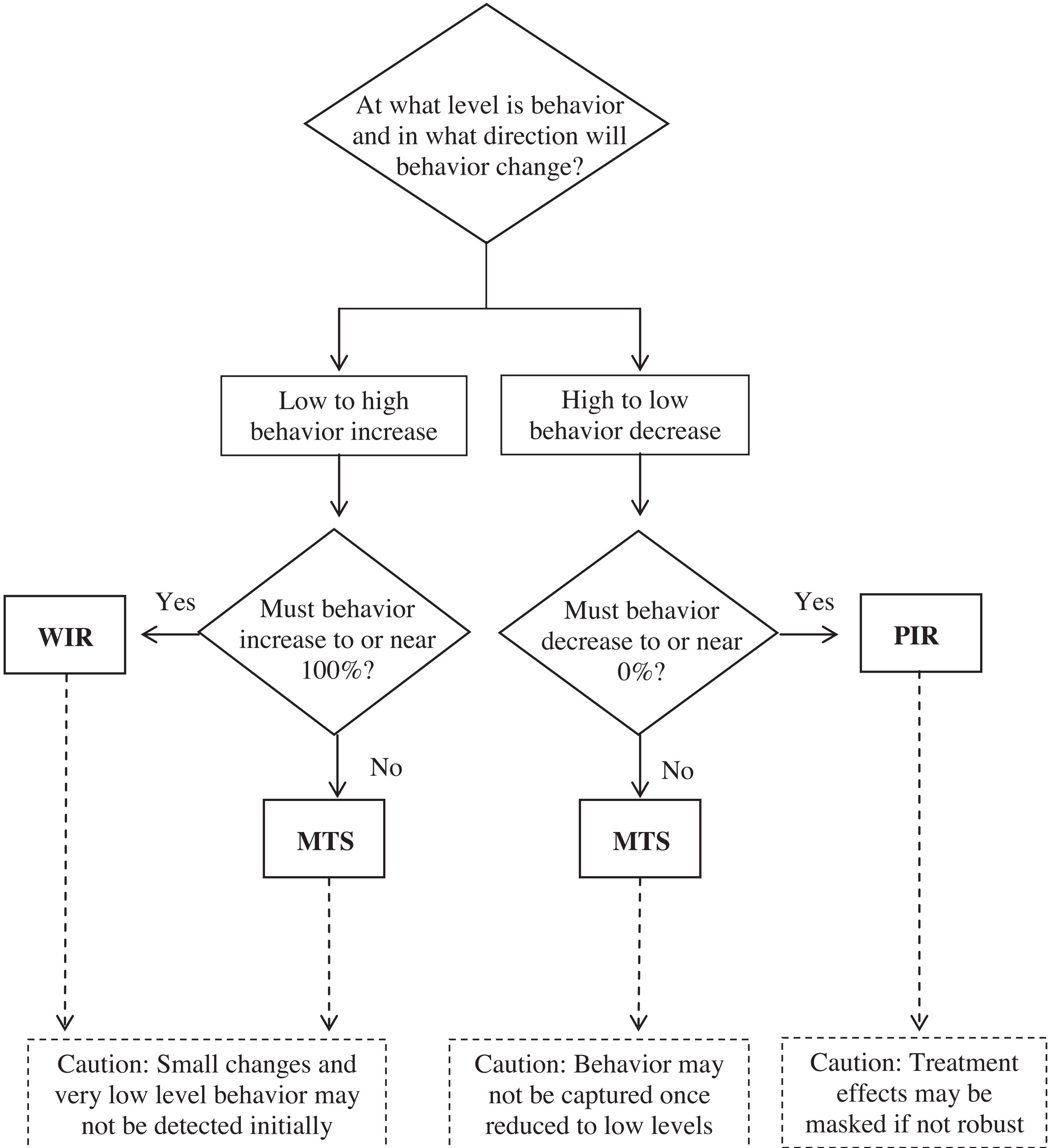
At what level is behavior and in what direction will behavior change?
Low to high behavior increase
High to low behavior decrease
Must behavior increase to or near 100%?
Must behavior decrease to or near 0%?
Yes
WIR
No
MTS
Yes
PIR
No
MTS
Caution: Small changes and very low level behavior may not be detected initially
Caution: Behavior may not be captured once reduced to low levels
Caution: Treatment effects may be masked if not robust

Figure 1

than would typically be considered an "effective" treatment for problem behavior (particularly when recoded with PIR). Rapp et al. (2007; 2008) reported similar findings.

Clinicians should be aware of the possibility that true behavior change may not be detected by the measurement system, dependent upon the known weaknesses of PIR and MTS. For example, if no behavior change is evident at first when monitoring using PIR or MTS, practitioners should determine whether the intervention was initiated for a behavior occurring at high levels and monitored using PIR or at low levels and monitored by MTS. In these cases, the data may not capture small changes in behavior, and practitioners should consider monitoring behavior over a longer period of time. Alternatively, the practitioner may decide that, if the behavior change is so small that it is unable to be captured by the data collection method, then the intervention should be altered to increase the significance or rate of behavior change.

In some cases, a measurement system may indicate a behavior change has occurred when little important change has taken place. Measurement also may suggest that the level of behavior has changed when in fact another dimension (e.g., distribution of the behavior over time) has changed. For example, in a scenario in which a student having long, infrequent tantrums begins to have short, frequent tantrums, the duration of tantrum has not changed but the distribution of behavior has. This type of problem, when a change in level is indicated by the data but has not actually occurred, is more frequently associated with PIR than with MTS (Rapp, 2008). Practitioners should be mindful of this possibility when interpreting data generated via PIR.

Summary

Selecting a measurement system is a pivotal step in effectively designing and evaluating behavioral intervention. Conclusions about the need and appropriateness of treatment goals and interventions rely on sound behavioral measurement that accurately reflects the behavior targeted for change. A practitioner must consider the potential impact of several factors in each clinical scenario, including characteristics of behavior as well as characteristics of the environment in which data are to be collected. Based upon the available literature, we provided a preliminary guide (see Figure 1) to help practitioners select a measurement system according to their specific case, while highlighting particular caveats and cautions. Again, we make the assumption that, when using the guide, the practitioner has already decided a discontinuous data system is most appropriate for their case.

A fair amount of research suggests treatment decisions can be meaningfully informed by data despite measurement error, but we make the case that practitioners must be cognizant of the characteristic patterns of measurement error that exist with discontinuous measurement. Practitioners should consider conditions under which such error may have particular significance in a given case. Continued research is necessary to further inform the data collection decisions made by practitioners in applied settings.

Author note

Correspondence regarding this article can be sent to Kate Fiske. Email: kfiske@rci.rutgers.edu.

References

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91–97.

Ciotti Gardenier, N., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities, 25,* 99–118.

Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions, 26*, 103–124.

Hanley, G. P., Cammilleri, A. P., Tiger, J. H., & Ingvarsson, E. T. (2007). A method for describing preschoolers' activity preferences. *Journal of Applied Behavior Analysis, 40*, 603–618.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis, 19*, 73–77.

Johnston, J. M., & Pennypacker, H. S. (2009). Observing and recording. *Strategies and Tactics of Behavioral Research 3rd Edition* (pp. 115–138). New York: Routledge.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis, 40*, 501–514.

Mudford, O. C., & Beale, I. L. (1990). The representativeness of observational samples of different durations. *Journal of Applied Behavior Analysis, 23*, 323–331.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis, 42,* 165–169.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis, 8,* 463–469.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis, 10,* 325–332.

Rapp, J. T., Colby, A. M., Vollmer, T., Roan, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions, 22,* 319–345.

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, D. N. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions, 23,* 237–269.

Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis, 4,* 533–537.

Action Editor: Jeffrey Tiger